Statistical Methods

Class Prediction Analyses for the BRCA1+ vs. BRCA1- and BRCA2+ vs. BRCA2- Classifications Shown in Table 2

Michael D. Radmacher
Richard Simon

mdradmac@helix.nih.gov
Rsimon@nih.gov

Classification of each tumor sample into one of two classes (e.g., BRCA1 mutation positive or negative) based on gene expression data was performed using a compound covariate predictor. The predictor is built in two steps. First, a standard two-sample *t*-test is performed to identify genes with significant differences (at level α) in log-expression ratios between the two tumor classes. Second, the log-expression ratios of differentially expressed genes are combined into a single compound covariate (REF: J.W. Tukey, Tightening the Clinical Trial, *Controlled Clinical Trials*, 14:266-285, 1993) for each tumor sample; the compound covariate is used as the basis for class prediction. The compound covariate for tumor sample *i* is defined as

$$c_i = \sum_j t_j x_{ij},$$

where t_j is the *t*-statistic for the two group comparison of classes with respect to gene *j*, x_{ij} is the log-ratio measured in tumor sample *i* for gene *j* and the sum is over all differentially expressed genes.

Cross-validated class prediction was performed using compound covariates. First, a tumor sample to be classified was removed from the data set. The remaining tumor samples (comprising the training set) were used to determine the differentially expressed genes between the two tumor classes. Using the log-expression ratios of these genes, the value of the compound covariate was computed for every tumor sample in the training set and a classification threshold was calculated (we used the midpoint of the means of the compound covariates for the two classes as the threshold). The class of the left out tumor sample was then predicted by computing the value of the compound covariate for the sample and determining which side of the threshold it fell on (i.e., which class mean it was closest to). The entire process was repeated so that every tumor sample was left out one time and its class membership predicted; the number of misclassified samples was tallied.

To determine whether the accuracy for predicting membership of tumor samples into given classes (as measured by the number of correct classifications) was better than the accuracy that could be attained for predicting membership into random groupings of the tumor samples, we examined the distribution of the number of misclassifications for data sets in which the class labels were permuted. We created 1000 random data sets by permuting class labels among the tumor samples. Cross-validated class prediction was performed on the resulting data sets as described above and the percentage of permutations that resulted in as few or fewer misclassifications as for the original labeling of samples was reported. If less than 5 % of the permutations resulted in as few or fewer misclassifications, the accuracy of prediction into the given classes (e.g., BRCA1 mutation positive or negative) was considered significant.